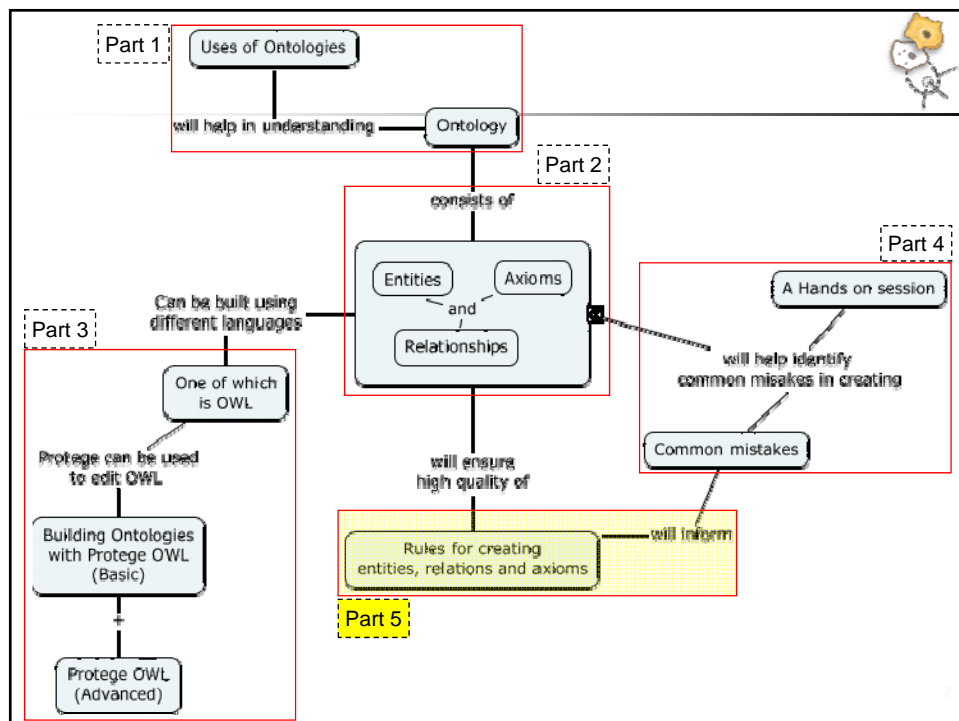


Summary and Take Home

(ICF Ontology and Protégé workshop 2008)

Nigam Shah

Stanford University
nigam@stanford.edu





Do's and Don'ts while creating your own ontology

Nigam Shah and Barry Smith



Need for a sound methodology

1. Ontologies must be intelligible both to humans (for annotation) and to machines (for reasoning and error-checking)
2. Unintuitive rules for classification lead to entry errors (problematic links)
3. Facilitate training of curators
4. Enable mapping with other ontology and terminology systems. Or avoid the need for mappings by having one ontology for each domain
5. Enhance harvesting of content through automatic reasoning systems

Need for a sound methodology



6. Ensure that lessons learned in building ontologies in the past are applied when building ontologies in the future
7. Knowing which ontology is already good enough to use for a given domain helps to avert silos
8. If the same set of evolving principles is being used by ontologists in different domains, then ontology building becomes a cumulative skill
9. Guidance works in every other area of science

5

First Commandment: Univocity



- Terms (including those describing relations) should have the **same meaning on every occasion** of use.
- In other words, they should refer to the same kinds of entities in reality
- Problem example: 'chromosome' in Sequence Ontology and in Cell Component Ontology means different things

6

Example of univocity problem



(Old) Gene Ontology:

- 'part_of' = 'may be part of'
 - flagellum part_of cell
- 'part_of' = 'is at times part of'
 - replication fork part_of the nucleoplasm
- 'part_of' = 'is included as a sub-list in'

7

Second Commandment: Positivity



- Complements of classes are not themselves classes.
- Terms such as 'non-mammal' or 'non-membrane' do not designate genuine classes.

8

Third Commandment: Objectivity



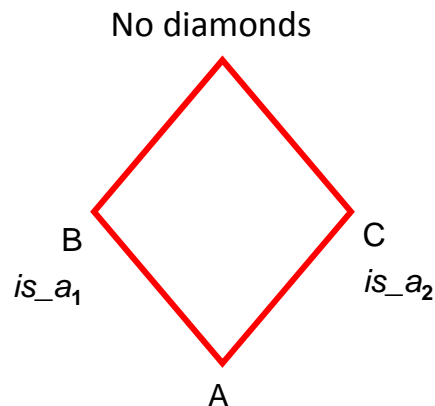
- Which classes exist is not a function of our biological knowledge.
- Terms such as 'unknown' or 'unclassified' or 'unlocalized':
 - do not designate biological natural kinds
 - do not designate differentiating characteristics [differentia] of biological natural kinds

9

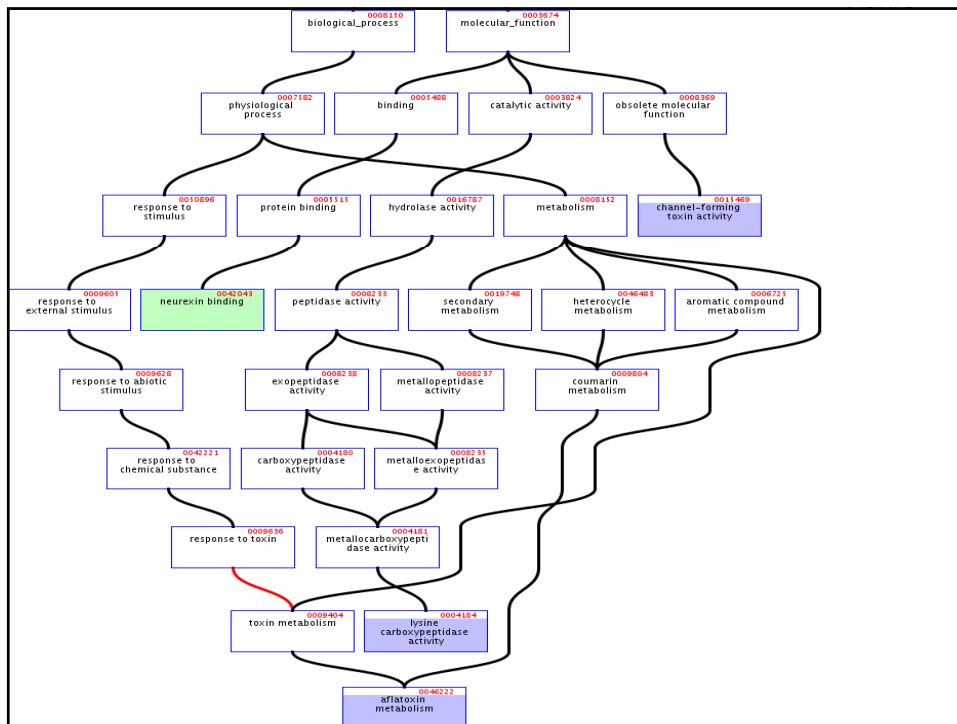
Fourth Commandment: Single Inheritance



No class in the asserted hierarchy should have more than one *is_a* parent on the immediate higher level



10



Fifth Commandment: Intelligibility of Definitions



- The terms used in a definition should be simpler (more intelligible) than the term to be defined
- otherwise the definition provides no assistance
 - to human understanding
 - for machine processing

Sixth Commandment: Basis in Reality



- When building or maintaining an ontology, always think carefully at how classes (types, kinds, species) relate to instances in reality
- If the Ontology is built to represent things that exist then the exchange format, data-model, xsd etc (application ontology), based on it always remains valid
 - ... even if our interpretation changes (B.P. – hypertension)

34

Seventh Commandment: Distinguish Universals and Instances



- A good ontology must distinguish clearly between
 - **universals (types, kinds, classes)**
 - and
 - **instances (tokens, individuals, particulars)**

35

The Seven Commandments



1. **Univocity:** Terms should have the same meanings on every occasion of use
2. **Positivity:** Terms such as 'non-mammal' or 'non-membrane' do not designate genuine classes.
3. **Objectivity:** Terms such as 'unknown' or 'unclassified' or 'unlocalized' do not designate biological natural kinds.
4. **Single Inheritance:** No class in a classification hierarchy should have more than one is_a parent on the immediate higher level
5. **Intelligibility of Definitions:** The terms used in a definition should be simpler (more intelligible) than the term to be defined
6. **Basis in Reality:** When building or maintaining an ontology, always think carefully at how classes relate to instances in reality
7. **Distinguish Universals and Instances**

36

Not everyone is a believer



- The world of biomedical research is a world of difficult trade-offs
- The benefits of formal (logical and ontological) rigor need to be balanced
 - Against the **constraints of computer tractability**,
 - Against the **needs of biomedical practitioners**.
- **WE CLAIM THAT: alignment and integration of biomedical information resources will be achieved only to the degree that these principles of classification and definition are followed**

37

Definitions should be intelligible to both machines and humans



- Machines can cope with the full formal representation
- Humans need to use modularity
- **Plasma membrane**
 - *is a cell part* [immediate parent]
 - that *surrounds* the **cytoplasm** [differentia]

18

Principle of Compositionality



- The meanings of compound terms should be determined by
 - the meanings of component terms
 - together with the rules governing syntax

19

Principle of Syntactic Separateness



- Do not confuse sentences with ontology terms
- If you want to say: No As are Bs
 - do not invent a new class of non-Bs and say A **is_a** non-B

20

Keep Epistemology Separate



- If you want to say that we do not know where As are located do not invent a new class of A's with unknown locations
 - Example: Holliday junction helicase complex **is-a** unlocalized
- A well-constructed ontology should grow linearly [monotonically];
 - it should not need to delete classes or relations because of increases in knowledge

21

Some other rules of thumb



1. Don't confuse entities with concepts
2. Don't confuse entities with ways of getting to know entities
 - a brain is not the same as its CT-scan
3. Don't confuse entities with ways of talking about entities
 - A person's medical record is not == person himself
4. Don't confuse entities with artifacts of your database representation ...
 - e.g. multiple dosing event in PharmGKB
5. An ontology should not change when the ontology language changes
 - The process of driving a car doesn't change whether you describe it in English or Spanish.

22

Guidelines for instances



- Every class has at least one instance
- Each child class has a smaller set of instances than its parent class
- Distinct classes on the same level never share instances
- Distinct leaf classes within a classification never share instances

23



Principles for Relations in Ontologies

Nigam Shah and Barry Smith



Ontologically rigorous relations

- Move from associative relations between meanings/concepts to strictly defined [ontological] relations between the entities themselves.
- It is not enough to consider just classes or types.
 - We need also to take account of *instances* and *time*
- The relations can then be used computationally

25

Benefits of well-defined relationships



- If the relations in an ontology are well-defined [All-Some structure], then reasoning can cascade from one relational assertion ($A R_1 B$) to the next ($B R_2 C$).
- Relations used in ontologies thus far have not been well defined in this sense.
- *Find all DNA binding proteins* should also find all transcription factor proteins because
 - *Transcription factor is_a DNA binding protein*

26

An unclear definition of is_a



- 'A' is more specific in meaning than 'B'
- HL7-RIM:
 - Individual Allele is_a Act of Observation
 - cancer documentation is_a cancer
 - disease prevention is_a disease

27

How to define the *is_a* relation



- What does *A is_a B* mean?
 - (A and B are types)
- For all *x*, if *x* **instance_of** *A* then *x* **instance_of** some *B*
- *cell division is_a biological process*

ALL-SOME STRUCTURE

28

An unclear definition of *part_of*



A part_of B:

A composes (with one or more other physical units) some larger whole *B*

This confuses relations between meanings or concepts with relations entities in reality

29

How to define A *part_of* B



- What does A ***part_of*** B mean?
- For all x, if x **instance_of** A then there is some y, y **instance_of** B and x **part_of** y
 - where '**part_of**' is the instance-level part relation
- *cell nucleus part_of cell*

ALL-SOME STRUCTURE

30

Kinds of relations



- Between classes:
 - *is_a, part_of, ...*
- Between an instance and a class
 - `this explosion` **instance_of** the class `explosion`
- Between instances:
 - Mary's heart **part_of** Mary

31

How many relations do we need?



Properties of Relations

1. Transitivity
 2. Symmetry
 3. Reflexivity
 4. Anti-Symmetry
 5. ...
- Avoid putting ‘_’ between arbitrary characters and calling it a relation
 - **is_somewhat_related_to** is the worst kind of relation to create!

32

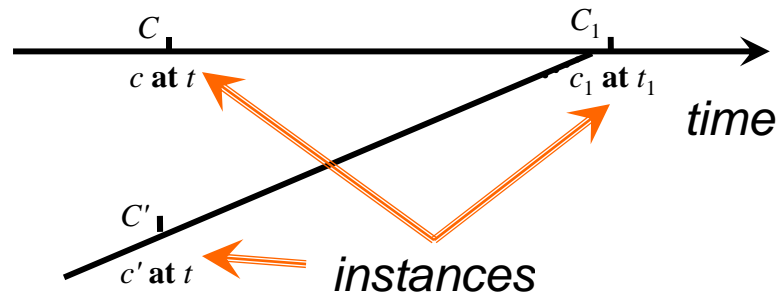
Don't forget instances when defining relations



- *part_of* as a relation between classes versus ***part_of*** as a relation between instances
 - *nucleus part_of cell*
 - your heart ***part_of*** you
- What holds on the level of instances may not hold on the level of universals
 - *nucleus adjacent_to cytoplasm*
 - **Not:** *cytoplasm adjacent_to nucleus*
 - *seminal vesicle adjacent_to urinary bladder*
 - **Not:** *urinary bladder adjacent_to seminal vesicle*

33

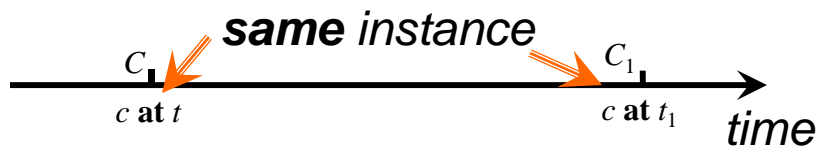
Time matters ... e.g. derives_from



zygote derives_from ovum
sperm

34

transformation_of



pre-RNA → mature RNA
child → adult

35



When representing 'Processes'

Nigam Shah



Inspired by BFO

- Keep a strict separation between things that exist continually and processes that unfold in time.
 - The hierarchies for continuants and occurrents should be disjoint.
- Be aware of where the process happens, what entities participate and for how what duration.
 - If you can not represent time, at least be explicit about the "location" and "participants" of the process

The “take home”



- Follow a methodology which enforces **clear, coherent definitions for entities and relationships**
- This promotes quality assurance
 - intent is not hard-coded into software
 - meaning of relationships is defined, not inferred
- **Enables automated reasoning** across ontologies and across data at different granularities

38



Common mistakes

Too much trust in natural language



- Too much trust in natural language leads to ambiguities. E.g. 'ontology' is used systematically ambiguous in natural language in order to refer:
 - (a) to a field of scientific research and
 - (b) a type of certain artifacts that are created by researchers.
 -
- These are quite different entities that have to be treated as distinct entities.
- People tend to trust natural language naively and assume the following correspondence:
 - One natural language expression corresponds to one entity.

40

Naive conceptualizations



- Many users embrace naive conceptualization, they declare things like
 - 'Fake Diamond is_a Diamond'
 - 'Absent leg is_a leg'.
 - Besides the fact that it is nonsense, this is wrong, because now 'Absent leg' will inherit all properties from 'leg'.

41

Logical ambiguity



Different readings of "part_of"

- cell nucleus part_of cell
 - **all** Xs are part of **some** Ys

All-Some STRUCTURE

- carrot part_of vomitus.
 - **some** Xs are part of **some** Ys

Some-Some STRUCTURE

42

Confusion caused by "is_a"



"is_a" used for both **instance_of** and **subtype**

- Correct: red is_a color, dictionary is_a book
- Incorrect: this flower is_a red, this dictionary is_a book
- Correct: the color of this book instance_of red

43

Inheritance



- We use `is_a` for inheritance. All properties of the parent node should be inherited by the child node: everything which holds of color holds of red.
- `part_of` does not support inheritance:
- not everything which holds of cell holds of cell nucleus
- something similar to inheritance holds for `instance_of`

44

Too much information in one ontology



- Most ontologies are `is_a` hierarchies of substance types. (Examples are the taxonomy of biological species or anatomical ontologies.)
- People often make the mistake to include relevant information in the ontology that belongs to another ontology, e.g. information about development state or pathology

Correct: animal, mammal, dog

Incorrect: animal, dog, brown dog, 6 year old brown dog

- The right solution is to keep the ontology of substance particulars and the ontology of attributes distinct.

45

ICD10 (1999): 587 codes for such accidents



•**V31.22 Occupant of three-wheeled motor vehicle injured in collision with pedal cycle, person on outside of vehicle, nontraffic accident, while working for income**

•W65.40 Drowning and submersion while in bath-tub, street and highway, while engaged in sports activity

•X35.44 Victim of volcanic eruption, street and highway, while resting, sleeping, eating or engaging in other vital activities

46

Acknowledgements



- NCBO is funded by NIH Roadmap initiative
- Protégé and Protégé-OWL are supported by grants and contracts from the NIH
- Daniel Rubin and Andrew Spear for contributing to slides and handout.

47



End